ChatR1: Reinforcement Learning for Conversational Reasoning and Retrieval Augmented Question Answering

Simon Lupart
University of Amsterdam
Amsterdam, The Netherlands
s.c.lupart@uva.nl

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Evangelos Kanoulas
University of Amsterdam
Amsterdam, The Netherlands
e.kanoulas@uva.nl

Abstract

We present ChatR1, a reasoning framework based on reinforcement learning (RL) for conversational question answering (CQA). Reasoning plays an important role in CQA, where user intent evolves across dialogue turns, and utterances are often underspecified, requiring contextual interpretation, query reformulation, and dynamic coordination between retrieval and generation. Unlike static 'rewrite, retrieve, and generate' pipelines, ChatR1 interleaves search and reasoning across turns, enabling exploratory and adaptive behaviors learned through RL. To address the challenge of sparse and delayed rewards in RL, we propose an intent-aware reward that provides turn-level feedback by aligning retrieval and reasoning with evolving user goals. ChatR1 demonstrates strong performance on both 3B and 7B model backbones, outperforming competitive models on five CQA datasets, measured by different metrics (F1, BERTScore, and LLM-as-judge). We include a diverse set of CQA datasets to cover topic shifts, evolving intents, mixed-initiative dialogues, and multi-document grounding, testing ChatR1's performance from various aspects. Ablation studies confirm the effectiveness of the intent-aware reward. Our analyses further reveal diverse reasoning trajectories and effective use of the search tool. ChatR1 also generalizes robustly across domains, demonstrating that RL-based reasoning enables more flexible and context-aware behavior than static CQA pipelines.[1]

1 Introduction

Reasoning models for information seeking have made rapid progress with the development of large language models (LLMs) and reinforcement learning (RL) (Guo et al., 2025). These systems are increasingly capable of fulfilling complex information needs, requiring multi-step reasoning, query decomposition, and tool use (Jin et al., 2025b; Chen et al., 2025; Song et al., 2025). By integrating search engines as external tools, RL-trained models move beyond static, supervised fine-tuned retrieval-augmented pipelines, instead, learning dynamic retrieval and reasoning from delayed rewards (Kaelbling et al., 1996; Chu et al., 2025). However, current RL reasoning frameworks remain limited to single-turn interactions with users, assuming explicit and isolated user questions (Li et al., 2025), whereas commercial systems are increasingly moving toward multi-turn conversational search.[2] Extending RL reasoning models to conversational scenarios introduces new reasoning complexities, as user intent evolves across turns and must be inferred from dialogue context (Dalton et al., 2022; Mo et al., 2025b).

Reasoning is particularly valuable for conversational question answering (CQA) as dialogues contain evolving and underspecified information needs that require reference resolution, query reformulation, and coordination between retrieval and answer generation (Elgohary et al., 2019). As illustrated in Figure 1, user utterances depend on earlier context ("one" referring to "European countries," "wind" to "wind energy," along with a comparison of previously cited countries). Systems must interpret context as an additional reasoning step (Radlinski and Craswell, 2017). CQA also spans heterogeneous datasets that capture distinct user interaction patterns, ranging from topic transitions and mixed-initiative to task-oriented exchanges (Wu et al., 2023; Adlakha et al., 2022). This diversity poses a unique generalization challenge, as models trained in one conversational domain often struggle to adapt to others (Mo et al., 2025b). As discussed by Chu et al. (2025), while supervised fine-tuning (SFT) captures behaviors from demonstration data,

[1] Models and datasets available on huggingface at `https://huggingface.co/collections/slupart/chatr1`

[2] e.g., Perplexity.ai: `https://www.perplexity.ai/`, OpenAI: `https://chatgpt.com/features/deep-research/`.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2257–2274

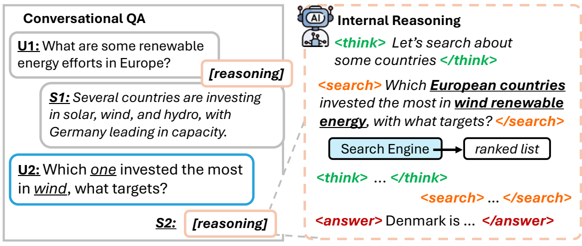

Figure 1: Reasoning in CQA requires the system to understand the user intent. The system resolves anaphoras and ellipses in the last user utterance (**U2**) during its internal reasoning, and formulates a search query, before producing the final answer.

reasoning from RL would enable broader generalization to conversational scenarios not explicitly seen during training. Despite the benefits of reasoning models for CQA, existing work relies exclusively on SFT (Acikgoz et al., 2025; Mo et al., 2025a; Xu et al., 2024), leaving RL-based conversational reasoning largely unexplored.

In this work, we propose an *RL framework for reasoning in CQA*, enabling models to learn retrieval and grounding strategies dynamically from interaction feedback. This approach differs from static retrieval pipelines by optimizing end-to-end reasoning behavior across multi-turn conversations.

While RL offers a promising framework for adaptive reasoning, its application to multi-turn CQA remains challenging because outcome-based rewards provide only delayed feedback with little guidance on which intermediate decisions, such as context interpretation or query formulation, contribute to success (Sutton and Barto, 2018). This reward sparsity amplifies instability and hinders learning in conversational settings, where reasoning steps are interdependent and user intent evolves continuously (Li et al., 2025; Zhang et al., 2025). We thus introduce an *intent-aware reward* that provides turn-level supervision by leveraging canonical user intent annotations from CQA datasets. In our experiments, we compare this signal with an alternative retrieval-based reward (e.g., document coverage), showing that our intent-aware shaping better aligns intermediate decisions with evolving goals and complements final rewards.

Our results show that RL alone offers limited benefits over SFT; however, introducing our intent-aware reward enables adapting RL reasoning to CQA. With this design, our proposed model, **ChatR1**, achieves substantial performance gains in both in-domain and out-of-domain settings, surpassing SFT baselines and highlighting the importance of reward shaping for multi-turn reasoning. Our main contributions are as follows:

- We introduce ChatR1, an RL-based reasoning model for CQA. ChatR1 optimizes multi-turn retrieval and generation end-to-end, learning dynamic behavior rather than a static pipeline. Extensive experiments demonstrate the performance gains and generalization capabilities of ChatR1 across different conversational complexities.
- We propose an *intent-aware reward* tailored for CQA, which reduces the sparsity of the reward signal by aligning retrieval behavior with evolving user intent across turns. Ablations show the benefit of this reward compared to other intermediate rewards.
- Our analysis further reveals that ChatR1 generates diverse reasoning paths, reflected in their lengths, and generalizes robustly across conversational domains. We also evaluate the performance of the retrieval and how it impacts generation performance to better understand the search usage of ChatR1.

2 Related Work

Conversational question answering is the task of responding to user queries within multi-turn dialogues, with dependence on previous turns and often grounding from external knowledge (Qu et al., 2020). Retrieval-augmented generation (RAG) has proven particularly effective in CQA, reducing hallucinations and providing factual grounding in answers (Mo et al., 2025a; Liu et al., 2025). Proprietary systems such as Perplexity Sonar, Gemini Grounding, GPT Search and Claude have also demonstrated promising search and generation capabilities (Miroyan et al., 2025). However, these industrial models remain closed-source, and it is unclear whether their behavior results from explicit reasoning or large-scale supervised pipelines composed of multiple static components. While academic research has advanced CQA with retrieval and generation, most approaches still rely on static retrieval pipelines and lack explicit reasoning mechanisms for deciding when and how to search (Xu et al., 2024). Similarly, in conversational search (CS), community evaluations emphasize user intent understanding and more complex user interactions (Dalton et al., 2020; Aliannejadi et al., 2024;

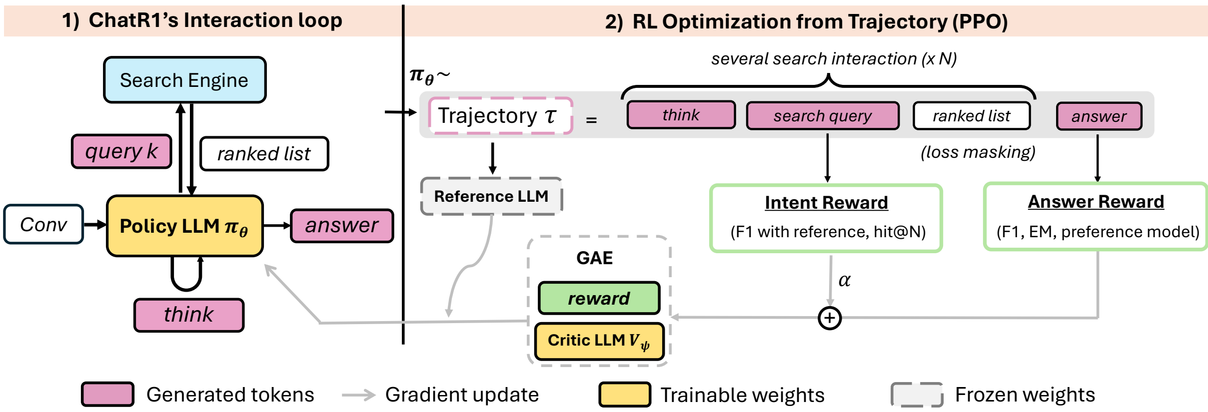

Figure 2: Overview of ChatR1's interaction loop: the policy model interleaves reasoning, search, and answering while interacting with a search environment. (Right) Full trajectories are sampled from the policy LLM and consist of all reasoning, search, and answer steps, with rewards assigned to search queries (intent-aware reward) and the final answer. These signals jointly train the policy under PPO optimization.

Abbasiantaeb et al., 2025; Gohsen et al., 2025). A common strategy is to model user information needs explicitly with *query rewriting*, to align underspecified user questions with user intent (Yu et al., 2021; Lupart et al., 2025a,b; Abbasiantaeb et al., 2026). RL has also been applied to query rewriting (Wu et al., 2021; Zhu et al., 2025); both works design rewards over retrieved passages to improve query rewriting. However, these approaches do not extend to answer generation, leaving reasoning in CQA largely unexplored. We aim to fill this gap with a reasoning model that can understand user intent and interact with both the user and a search engine.

Reasoning and QA with RL. Recent work has applied RL to train LLMs as reasoning agents for information seeking (Zhang et al., 2025; Li et al., 2025). Approaches such as Search-R1 (Jin et al., 2025b), R1-Searcher (Song et al., 2025), ReSearch (Chen et al., 2025), DeepResearcher (Zheng et al., 2025) optimize policies that decompose complex questions, issue multiple queries, and integrate evidence, achieving strong results on knowledge-intensive QA. While early studies (Jin et al., 2025b) relied solely on sparse trajectory-level rewards, some later work such as SearchR1++ (Jin et al., 2025a) and StepSearch (Wang et al., 2025) focused on the credit assignment problem and reward sparsity by introducing step-level rewards on intermediate reasoning or retrieval quality. In contrast, our approach extends this line of work by defining an intermediate reward tailored to CQA, capturing user intent at each dialogue turn to provide finer-grained learning signals.

In parallel, advances in agentic tool use have complemented RL-based reasoning, enabling models to plan and invoke external APIs or search engines (Schick et al., 2023; Singh et al., 2025). Among these, CALM (Acikgoz et al., 2025) extends reasoning to multi-turn dialogues, interleaving tool calls with user interactions while emphasizing the challenge of maintaining coherent user intent across turns. However, CALM remains trained through SFT, relying on demonstration data; extending it to RL could enable better exploration of multi-turn reasoning strategies.

3 Methodology

In this section, we introduce the CQA notation before presenting ChatR1, the RL objective, and our intent reward mechanism.

3.1 Notation and Problem Setup

We consider a dataset $\mathcal{D}$ of user–system conversations, each composed of multiple turns, and a collection of passages $\mathcal{C}$. At each turn, the system receives the conversation history $\mathcal{H}$ (all previous utterances) and the current user query q. The CQA task is to generate an answer y to q, leveraging context from $\mathcal{H}$ and grounding in $\mathcal{C}$. We further define the user intent via rewritten queries q^{rw}, obtained from human annotations, that resolve contextual references and ambiguities in q. These rewrites are available only during training as supervision.

3.2 ChatR1 RL Objective

Figure 2 illustrates the main components of ChatR1, along with the trajectory and reward structure. ChatR1 is a policy model π_θ that, at each turn, generates a trajectory τ. The trajectory consists of a reasoning trace including thinking, intermediate search queries $Q = \{q^k\}_{k=1}^K$ issued to the search engine $\mathcal{R}$, and retrieved passages, followed by the final answer y. The policy model's instructions define all special tokens linked to the external tools. In particular, the <search> token triggers the retriever, and the retrieved documents are then added to the trajectory. The full instruction of ChatR1 is provided in Table 1.

Objective. The objective of ChatR1 is to maximize the expected reward given the conversation history, the last user utterance, and the search engine, while regularizing the policy toward the reference policy:

$$\mathcal{J}(\theta) = \mathbb{E}_{(q,\mathcal{H})\sim\mathcal{D},\ \tau\sim\pi_\theta(\cdot|q,\mathcal{H};\mathcal{R})} R(\tau) - \beta\, D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big)\ . \tag{1}$$

Policy optimization. We optimize this objective using Proximal Policy Optimization (PPO) (Schulman et al., 2017), a policy-gradient algorithm designed to stabilize updates. PPO maximizes the clipped surrogate objective:

$$\mathcal{L}^{\mathrm{PPO}}(\theta) = \mathbb{E}_{(q,\mathcal{H};\mathcal{R};i)\sim\mu} \Big[\min\big(\rho_i(\theta)\,\hat{A}_i,\ \mathrm{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_i\big)\Big]\ , \tag{2}$$

where the expectation under μ averages over conversation examples from the dataset, trajectories sampled from the policy, and generated token positions, following previous works on token loss masking (Jin et al., 2025b). We define the ratio ρ as the probability ratio between the new and old policies.

$$\rho_i(\theta) = \frac{\pi_\theta(\tau_i \mid \mathcal{H}, q, \tau_{<i}; \mathcal{R})}{\pi_{\theta_{\mathrm{old}}}(\tau_i \mid \mathcal{H}, q, \tau_{<i}; \mathcal{R})}\ . \tag{3}$$

We also represent $\hat{A}_i$, the estimated advantage at token position i, which quantifies how much better the chosen token τ_i performs compared to the expected value predicted by the critic.

$$\hat{A}_i = \delta_i + (\gamma\lambda)\,\delta_{i+1} + \cdots + (\gamma\lambda)^{T-i-1}\,\delta_{T-1}\ , \tag{4}$$

$$\delta_i = \gamma\, V_\psi(\tau_{i+1}) - V_\psi(\tau_i)\ , \tag{5}$$

with $V_\psi(\tau_T) = R(\tau)$ the final reward for the last token of the trajectory, and V_ψ the surrogate critic model. This ensures that rewards are propagated across all generated tokens, allowing the model to assign credit throughout the reasoning and answer generation process. V_ψ is optimized with Generalized Advantage Estimation (GAE) (Schulman et al., 2015) following existing literature.

```
You are a helpful assistant tasked with
answering a user query. Your primary goal is
to generate a complete and informative answer.
If the query is ambiguous or refers to
earlier context (e.g., pronouns or ellipsis),
use the conversation history provided below
to resolve it.
- Always perform your reasoning inside <think>
...</think>.
- If external information is needed, use
<search>your query</search>.
- Retrieved documents will appear between
<information>...</information>.
- You may issue multiple search queries if
needed.
- Once you have enough information, provide a
complete answer within <answer>...</answer>.

Conversation context: {context_block}
User query: {last_user_utterance}
```

Table 1: Instructions for the ChatR1 policy LLM.

3.3 Reward Modeling

We design a composite reward function that captures both the quality of the final answer and the understanding of user intent throughout the trajectory. For each trajectory τ, composed of a sequence of issued queries $Q = \{q^1, \ldots, q^K\}$ and final answer y, the total reward is defined as:

$$R(\tau) = R_{\text{answer}}(y) + \alpha\, R_{\text{intent}}(Q)\ , \tag{6}$$

where α balances the contribution of answer quality and intent understanding.

Answer reward. The answer reward R_{answer} evaluates how well the final system response aligns with the gold reference y^*. Following prior work (Jin et al., 2025b; Song et al., 2025), we adopt a word-level F1 score to capture semantic overlap:

$$R_{\text{answer}}(y) = \mathrm{F1}(y, y^*)\ , \tag{7}$$

where $\mathrm{F1}(a,b) = \frac{2\,\mathrm{P}(a,b)\,\mathrm{R}(a,b)}{\mathrm{P}(a,b)+\mathrm{R}(a,b)}$, computed from word-level precision and recall. This reward is assigned only at the end of the trajectory, encouraging coherent intermediate reasoning steps that ultimately improve the final answer.

Intent reward. The intent reward R_{intent} provides additional feedback for the quality of intermediate search queries. It measures how well any issued query q^k captures the user intent, expressed by the human-authored rewrite q^{rw}:

$$R_{\text{intent}}(Q) = \max_{q^k \in Q} \text{F1}(q^k, q^{rw}) . \quad (8)$$

Taking the maximum allows for exploration while ensuring that at least one search query aligns with the user intent. Compared to previous overlap-based query rewards such as StepSearcher (Wang et al., 2025), our proposed reward explicitly targets user-intent alignment in the CQA setting, offering a denser, retrieval-agnostic supervision signal that remains informative even when retrieval or final answering fails.

As an alternative reward, we also test a simpler retrieval-based variant that rewards retrieved relevant passages from the top-N of any of the issued search queries (Jin et al., 2025a).

4 Experiments

4.1 Experimental Setup

Dataset statistics. We trained and compared ChatR1 on five CQA datasets, containing multi-turn dialogues with evolving user intent. Table 2 lists the train and test splits, as well as the main challenges of each dataset. TopiOCQA (Adlakha et al., 2022) contains open-domain conversations where questions often involve topic shifts, requiring models to retrieve and reason over new information as the dialogue progresses. QReCC (Anantha et al., 2021) is a large-scale conversational dataset on a large web corpus, paired with reformulations and gold answers. INSCIT (Wu et al., 2023) focuses on information-seeking dialogues with mixed initiative (e.g., clarification questions). This requires models to understand user feedback and use it accordingly. FaithDial (Dziri et al., 2022) is a conversational QA dataset, as an improved version of WoW (Dinan et al., 2018), with better faithfulness and grounding to passages of the collection. Finally, MultiDoc2Dial (Feng et al., 2021) is a goal-oriented information-seeking dataset grounded in multiple governmental documents. More details can be found in Appendix B.

Evaluation metrics. We report the results of the generation task in terms of F1, BERTScore, and using LLM-as-judge, all with respect to a human gold reference answer. The prompt for the LLM-as-judge metric is provided in Appendix A, with `gpt-4o-mini`, following established practice (Song et al., 2025; Rau et al., 2024). For retrieval, we report standard IR metrics, including Normalized Discounted Cumulative Gain (nDCG), recall (R), and mean reciprocal rank (MRR). We refer to hit@N as the proportion of queries on which a relevant passage is within the top N of the retrieved list, following previous works (Järvelin and Kekäläinen, 2000; Abbasiantaeb et al., 2025).

Dataset	#Turns	Challenges
TopiOCQA (Adlakha et al., 2022)	45k 2.5k	Topic shifts, evolving intent
QReCC (Anantha et al., 2021)	63k 16k	Large-scale corpus, query reformulation
INSCIT (Wu et al., 2023)	1.8k 3.3k	Mixed initiatives, open intent
MultiDoc2Dial (Feng et al., 2021)	18k 3.3k	Multi-doc grounding, in-domain reasoning
FaithDial (Dziri et al., 2022)	18k 3.5k	Faithfulness, hallucination control

Table 2: Challenges of respective datasets with numbers of turns in train and test splits.

Baselines. We compare ChatR1 with a wide range of competitive baselines: zero-shot, supervised fine-tuned, and RL models. Direct inference (**DI**) with GPT-3.5, Claude, and Qwen2.5-3B-Instruct models generates answers from internal model knowledge, with the variant using chain-of-thought (**CoT**). **IRCoT** similarly uses chains of thought in combination with a search engine using the prompt of ChatR1 in zero-shot. Retrieval augmented generation (**RAG**) uses query rewrite from LLM to do conversational retrieval, and then uses the LLM to generate the final answer based on retrieved passages. All of these models take advantage of their long input context windows to understand the history and the last user utterance.

We further compare ChatR1 with supervised fine-tuned approaches, each trained on the same datasets as ChatR1. **CoT R1** is a chain-of-thought approach trained with an F1 RL-reward on the final answer, similar to ChatR1, without access to the retrieval tool. Supervised fine-tuning (**SFT**) is trained with a next-token prediction loss to generate the answer. **QR Search R1** uses query rewrite on the input conversation and then trains a Search R1 model on the query rewrites distribution. Finally, we report the results of three RAG state-of-the-art methods, **UniConv** (Mo et al., 2025a), **conv-ANCE +Mistral-7B** and **ChatRetriever +Mistral-7B** (Mao et al., 2024). Those three models are fine-tuned on a large conversational dataset consist-

Method	RAG	LLM	TopiOCQA		QReCC		INSCIT		MD2Dial		FaithDial	
			F1	**Bert**	**F1**	**Bert**	**F1**	**Bert**	**F1**	**Bert**	**F1**	**Bert**
Zero-Shot												
Qwen-Instr. (DI)	No	Qwen-3b	6.7	57.3	13.3	55.3	17.9	58.1	13.2	64.4	10.9	64.8
Qwen-Instr. (CoT)	No	Qwen-3b	6.7	58.6	12.8	58.5	16.4	62.4	10.5	63.6	9.6	64.8
IRCoT	RAG	Qwen-3b	8.9	61.0	13.1	55.6	20.4	67.3	13.3	67.5	9.6	64.7
Qwen-Instr. (RAG)	RAG	Qwen-3b	8.8	54.7	15.5	64.5	13.0	49.3	18.8	75.1	12.3	73.8
ChatGPT (DI)	No	GPT-3.5	25.5	77.5	22.6	75.6	22.8	81.1	21.6	81.7	12.9	80.9
Claude (DI)	No	Claude	27.2	–	25.0	–	27.0	–	–	–	–	–
Supervised Fine-tuning or RL Training												
conv-ANCE +Mis.	RAG	Mistral 7b	27.2	–	25.9	–	24.8	–	–	–	–	–
ChatRetriever +Mis.	RAG	Mistral 7b	28.3	–	26.3	–	30.3	–	–	–	–	–
UniConv	RAG	Mistral 7b	<u>29.6</u>	–	26.2	–	**33.2**	–	–	–	11.6	–
CoT R1	No	Qwen-3b	12.5	70.3	17.7	72.6	24.1	84.0	18.0	80.2	14.5	80.9
SFT	No	Qwen-3b	18.0	78.8	23.3	<u>80.0</u>	16.9	56.9	25.4	<u>84.2</u>	<u>18.6</u>	83.8
QR Search R1	RAG	Qwen-3b	20.1	72.7	20.4	79.6	27.5	84.0	23.1	82.1	14.4	82.2
ChatR1 (w/o $R_{\text{int.}}$)	RAG	Qwen-3b	24.4	73.0	27.0	78.5	31.3	84.4	<u>26.4</u>	77.4	15.5	81.3
ChatR1-3b	RAG	Qwen-3b	29.4†	**80.9**†	<u>28.0</u>†	79.2†	**33.2**†	**85.5**†	26.0	83.1†	**19.2**†	<u>84.0</u>†
ChatR1-7b	RAG	Qwen-7b	**30.6**‡	<u>79.5</u>	**31.0**‡	**80.7**‡	<u>32.8</u>	**85.5**	**31.2**‡	**84.5**‡	18.1	**84.8**‡

Table 3: Conversational response generation comparison of ChatR1 on five datasets, for zero-shot, SFT and RL baselines. Superscript † and ‡ are paired t-tests ($p < 0.05$) between w/ vs. w/o intent reward, and between 3B and 7B variants, respectively. Best results in bold, second best underlined.

ing of TopiOCQA and synthetic dialogs. They are also trained in two steps: first, the retrieval model with a contrastive loss, then the generation model with SFT. They also feature 7B encoder models for retrieval, while ChatR1 retrieval is only 300M.

Model and training details. We rely on `Qwen2.5-3B-Instruct` and its 7B variant as the base LLMs (Yang et al., 2025), fine-tuned on each five datasets. Retrieval model is `e5-base-v2` used in zero-shot. ChatR1 and other baselines use the top-3 retrieved passages; we also limit the number of search calls to two, following previous works (Jin et al., 2025b,a). Query rewrites are human annotations; when not available, we used GPT-4.1 to generate them. The batch size of the policy model is 512, with a PPO micro batch size of 64, a max prompt length of 3500 tokens, and a learning rate of 1e-6 for the actor model. We rely on the GAE algorithm for the critic optimization. Both policy and critic models are initialized with the same LLM, with weights fine-tuned independently. We train for 500 steps, with a saved checkpoint every 50 training steps. We use the final checkpoint unless training rewards collapse. Additional details are provided in Appendix D. All experiments are conducted on 4 H100 GPUs.

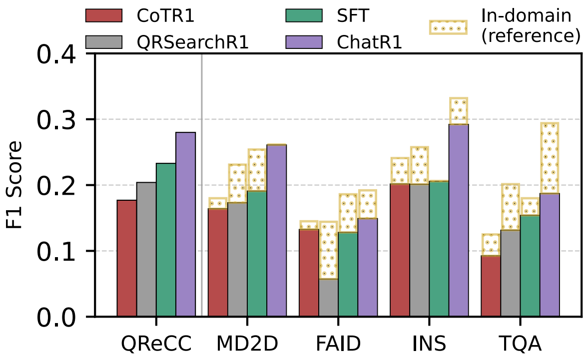

Figure 3: Out-of-domain transfer performance for models trained on QReCC, with in-domain reference.

4.2 Results

Conversational response generation. Table 3 presents the main results comparing ChatR1's generation performance with baselines. From the results, we observe that (i) ChatR1-3B achieves higher performance than closed-source commercial LLMs (ChatGPT, Claude) while using far fewer parameters, highlighting the effectiveness of RAG over parameter scaling. (ii) ChatR1-3B surpasses all SFT- and RL-based 3B models in both F1 and BERTScore across most datasets. Moreover, ChatR1-3B matches or exceeds 7B supervised RAG baselines such as UniConv and ChatRetriever+Mistral on TopiOCQA, QReCC, and INSCIT. Notably, while UniConv and ChatRetriever rely on 7B retrieval backbones, ChatR1-3B employs a 300M retriever, highlighting its strong ability to leverage retrieval tools effectively. (iii) We further observe that scaling ChatR1 from 3B to 7B parameters yields consistent performance gains,

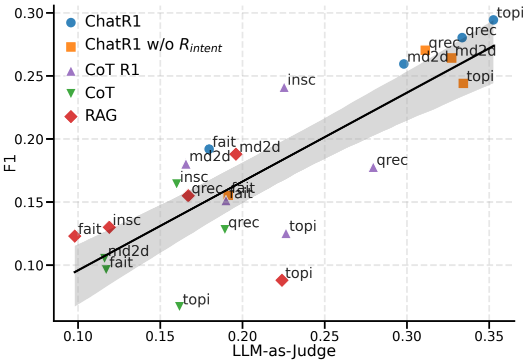

Figure 4: Correlation between using F1 score and LLM as judge on ChatR1 and baselines across datasets.

with an average improvement of 1.4 F1 points and 0.5 BERTScore. With this scaling, ChatR1-7B outperforms most compared baselines overall.

Generalization. To further assess the generalization ability of ChatR1, Figure 3 reports its performance in out-of-domain settings, where models trained on QReCC are evaluated on the four remaining datasets. We also highlight in gold the reference in-domain performance. Interestingly, ChatR1-3B shows minimal loss on MultiDoc2Dial (only 0.2) and generally lower loss than other baselines across the remaining datasets. ChatR1 shows slightly higher loss on TopiOCQA, likely due to the shorter answer lengths of the dataset responses, yet still surpasses all baselines. We also observe that ChatR1-3B still surpasses ChatGPT on three of the four datasets in zero-shot. This highlights how ChatR1 learns to use the retrieval tool, rather than overfitting to domain-specific datasets.

LLM evaluation. Figure 4 further illustrates ChatR1's performance when evaluated with LLM-as-judge. LLM-as-judge and F1 generally agree for in-domain evaluation with a Pearson's r of 0.83, indicating agreement between both metrics. The results also demonstrate the clear benefit of retrieval and fine-tuning. Finally, many zero-shot approaches, even though they have low F1, still have reasonable LLM-as-judge performance, showing that fine-tuning enables the model to also learn the distribution of words in the answer, and the targeted format of gold answers.

4.3 Analysis

While the previous section discussed ChatR1's performance, this section presents an ablation study and analyses of reasoning length and retrieval quality to examine variability in model outputs and

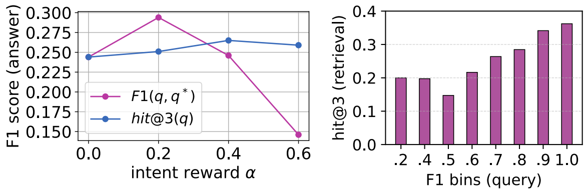

Figure 5: (Left) Analysis of intent rewards, contrasting passage-level rewards (hit@k) with query-level rewards (F1), and their impact on conversational QA response quality. (Right) Relationship between hit@k and query-level F1. Both on TopiOCQA with ChatR1-3B.

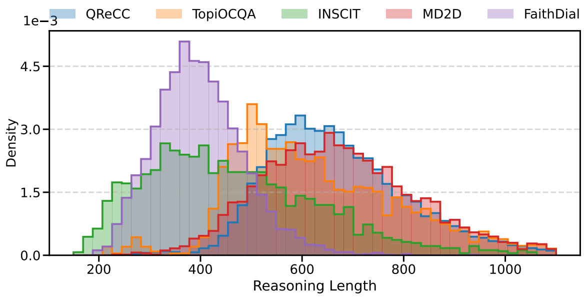

Figure 6: Reasoning length of ChatR1-7B when trained on the different datasets.

ChatR1-3B	TQA	QRC	INS	MD2D	FD
Jacc. overlap	0.52	0.25	0.12	0.29	0.57

Table 4: Jaccard overlap between consecutive search queries used by ChatR1.

retrieval behavior.

Reward ablations. We conduct ablations on the generation of ChatR1 to study the effect of the intent reward. First, as shown in Table 3, ChatR1-3B outperforms the base model without the intent reward by an average of 2.2 F1 points. We also explored the impact of the α parameters and reward in Figure 5. In particular, we observe that the intent reward on search queries directly $F1(q, q^*)$ leads to better performance compared to the reward on the retrieved passages with hit@3. This can be explained from several perspectives: (i) F1 on search queries is less sparse, providing a stronger learning signal for PPO and more stable optimization with dense rewards (Zhu et al., 2025; Sutton and Barto, 2018); (ii) it offers a direct signal on query formulation, independent of the search engine, thereby disentangling retrieval errors from formulation errors; and (iii) passage relevance judgments have annotation gaps (i.e., some passages may contain valid answers but are not labeled as relevant), causing hit-based rewards to be incomplete. In con-

Model	TopiOCQA		QReCC	
	N@3	**R@10**	**N@3**	**R@10**
Query Rewrite				
Gold Rewrite	30.6	52.5	40.6	67.2
QuReTeC	10.5	20.2	32.6	55.0
Qwen-3B QR	19.6	35.3	27.2	45.5
T5QR	22.2	37.6	31.8	53.1
LLM4CS	**26.7**	43.3	**42.1**	**66.4**
Encoder-based Retrieval				
ConvDR	26.4	43.5	35.7	58.2
Retrieval-as-a-Tool				
IRCoT-3B	18.7	35.6	31.2	52.8
ChatR1-3B	24.1	43.7	36.4	60.2
ChatR1-7B	**26.7**	**46.9**	37.0	61.1

Table 5: Retrieval-as-a-tool performance of ChatR1 on TopiOCQA and QReCC datasets with dense retrieval.

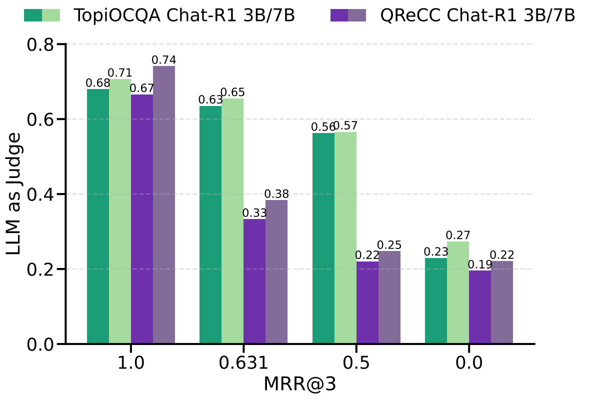

Figure 7: Retrieval quality effect on generation performance for ChatR1.

trast, query-level rewards directly assess the quality of the formulated query and are not affected by such labeling gaps. From a requirements perspective, constructing complete passage-query relevance pairs is considerably more expensive than generating rewrites. Finally, we observe that the best performance is achieved with a reward ratio of 0.2/1.0 between retrieval and generation rewards. As shown on the right side of the figure, $F1(q, q^*)$ is less sparse but still correlates with the hit metric.

Reasoning paths. Figure 6 illustrates the diversity in reasoning length across the five evaluated datasets. We observe that MultiDoc2Dial and QReCC exhibit the longest reasoning traces, while FaithDial features comparatively shorter ones. This is expected, as MultiDoc2Dial involves multi-hop CQA, requiring the model to perform deeper reasoning and multiple retrieval steps to synthesize relevant information. In contrast, INSCIT shows a much more dispersed distribution, which can be attributed to its mixed-initiative nature; some turns demand substantial reasoning due to user-driven complexity, while others remain relatively simple.

To further investigate whether the model tends to produce near-duplicate rewrites, we computed the Jaccard word-overlap between ChatR1's first and second queries when it generates > 2 queries for a turn. From Table 4, we observe some low similarity on INSCIT, MultiDoc2Dial, and QReCC with all values < 0.30, but higher similarity on FaithDial and TopiOCQA. This shows that the model explores the collection, which is the targeted behavior.

Conversational search. Table 5 highlights the retrieval performance of ChatR1, which leverages retrieval as a tool without relying on direct supervision. Despite this, ChatR1 achieves results that are on par with state-of-the-art query rewriting and encoder-based methods. In particular, ChatR1-7B matches or even surpasses several supervised baselines across both TopiOCQA and QReCC, demonstrating the effectiveness of reasoning-driven retrieval emerging from interactive learning. The consistent gains from 3B to 7B also show that scaling enhances retrieval reasoning rather than merely memorization. Overall, these findings confirm that ChatR1 can autonomously learn to perform retrieval effectively, approaching the performance of systems explicitly trained for query rewriting.

Figure 7 further illustrates how retrieval quality impacts generation performance in ChatR1. The model exhibits a clear trend where higher retrieval accuracy translates to improved answer quality. Both 3B and 7B variants show consistent improvements across retrieval bins, confirming that ChatR1 effectively adapts its reasoning process to the quality of retrieved evidence. The 7B model offers uniform gains across all bins, suggesting that scaling enhances overall robustness rather than compensating only for poor retrieval. Interestingly, the performance jump from low to mid retrieval quality is more pronounced on QReCC than on TopiOCQA. This is likely because QReCC, as a web-based dataset with longer passages, requires the model to process and verify more context within each retrieved document, making retrieval accuracy more critical to final generation quality.

PPO vs GRPO. While ChatR1 is trained with PPO, we provide in Figure 8 training curves of

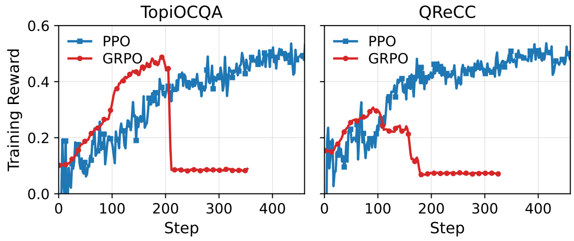

Figure 8: Training of ChatR1-3B with PPO and GRPO.

Method / Dataset	TQA	QRC	INS	MD2D	FD
ChatR1 +PPO	29.4	28.0	33.2	26.0	19.2
+GRPO	32.9	26.9	29.2	26.5	17.0

Table 6: Performance comparison between PPO and GRPO on ChatR1-3B across datasets.

ChatR1 with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), on both TopiOCQA and QReCC. Overall we observe that PPO is more stable than GRPO across datasets. ChatR1 trained with GRPO usually collapses after 100 or 200 steps, while PPO can be trained for longer. This aligns with previous work (Jin et al., 2025a) on the stability comparison of GRPO with PPO. Table 6 further reports the performance comparison of models trained with both optimizations. Despite the lower stability of GRPO, we observe that on two datasets, GRPO leads to better final performance.

Training intent quality. ChatR1 leverages human query rewriting as a user intent reward, to explicitly express the user's information need. While several datasets contain human rewrite, some such as TopiOCQA do not. Table 7 compares using pseudo intent/rewrite labels from T5, Mistral, and GPT-4.1 when training ChatR1. Overall, the performance follows the same pattern: better rewrites lead to higher F1 and BERTScore, while weaker signals hurt performance.

Retrieval tool. Table 8 compares ChatR1-3B and 7B under different retrieval setups at inference time. Models are trained with a dense retriever and evaluated with BM25, dense retrieval, and dense retrieval with reranking. We observe that BM25 leads to a clear drop in performance, especially for the 3B model. In contrast, adding reranking on top of dense retrieval consistently improves results across both datasets. Interestingly, the 7B model is also more robust with smaller degradation when switching to BM25 (-11.8 F1 for 3B while -7.8 F1 for 7B). We also see that models trained on QReCC have

	Training Signal		Response Generation	
Model	Intent/Rewrite	N@3	F1	BERTScore
ChatR1	T5	22.2	24.7	71.7
ChatR1	Mistral	26.0	27.7	77.6
ChatR1	GPT-4.1	30.6	29.4	80.9

Table 7: ChatR1 (3B) trained from silver quality intent reward, and impact on the downstream generation performance on TopiOCQA.

Retrieval Method	TopiOCQA	QReCC
ChatR1-3B +Dense	29.4	28.0
ChatR1-3B +BM25	17.6	26.6
ChatR1-3B +Dense-Rerank	31.8	29.9
ChatR1-7B +Dense	30.6	31.0
ChatR1-7B +BM25	22.8	29.4
ChatR1-7B +Dense-Rerank	32.1	33.1

Table 8: Generalization of ChatR1 to different retrieval systems during inference (F1-Score).

much lower drop from using BM25 on QReCC compared to TopiOCQA, which can be explained as the retrieval task on QReCC is easier than on TopiOCQA.

5 Conclusion

We presented ChatR1, an RL-based reasoning framework for CQA, where user intent evolves across turns and must be inferred from context. Unlike static SFT pipelines that separate rewriting, retrieval, and generation, ChatR1 jointly optimizes these steps end-to-end. To address the sparsity of the outcome reward, we introduced an intent-aware reward component that aligns retrieval and reasoning with evolving user intents. Our experiments show that this design significantly improves both in-domain and out-of-domain performance over SFT baselines, while our analysis reveals adaptive reasoning behaviors such as context-aware query reformulation and variable reasoning depth. Future work will explore dialogue-level optimization with simulated users and preference-based feedback, advancing RL-based reasoning for interactive information-seeking tasks.

Acknowledgements

This research was partly supported by the Swiss National Science Foundation (SNSF), grant number 215742, on Personality and Conversational Information Access (PACINO).

6 Limitations

While ChatR1 demonstrates that RL can improve reasoning in CQA, several aspects remain open for future work. Our framework relies on PPO/GRPO; exploring alternative optimization strategies, such as off-policy methods, or curriculum-based training could improve stability and sample efficiency. Current experiments focus on dialogues of moderate length (10–20 turns), whereas real-world interactions can be longer and require stronger memory and context modeling. Although smaller models already learn useful behaviors under RL, scaling to larger backbones may reveal stronger emergent reasoning and generalization. We also did not consider personalization or user-specific adaptation, which would be essential for more proactive and mixed-initiative dialogue. Finally, RL training introduces additional computational cost, both at training and inference; developing more efficient optimization schedules remains an important direction.

References

Zahra Abbasiantaeb, Simon Lupart, and Mohammad Aliannejadi. 2026. Generating multi-aspect queries for conversational search. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8201–8217, Rabat, Morocco. Association for Computational Linguistics.

Zahra Abbasiantaeb, Simon Lupart, Leif Azzopardi, Jeffrey Dalton, and Mohammad Aliannejadi. 2025. Conversational gold: Evaluating personalized conversational search system using gold nuggets. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3455–3465, New York, NY, USA. Association for Computing Machinery.

Emre Can Acikgoz, Jeremiah Greer, Akul Datta, Ze Yang, William Zeng, Oussama Elachqar, Emmanouil Koukoumidis, Dilek Hakkani-Tür, and Gokhan Tur. 2025. Can a single model master both multi-turn conversations and tool use? CoALM: A unified conversational agentic language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12390, Vienna, Austria. Association for Computational Linguistics.

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. Trec ikat 2023: A test collection for evaluating conversational and interactive knowledge assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 819–829, New York, NY, USA. Association for Computing Machinery.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, and 1 others. 2025. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Jeffrey Dalton, Sophie Fischer, Paul Owoicho, Filip Radlinski, Federico Rossetto, Johanne R. Trippas, and Hamed Zamani. 2022. Conversational information seeking: Theory and application. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3455–3458, New York, NY, USA. Association for Computing Machinery.

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1985–1988, New York, NY, USA. Association for Computing Machinery.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcel Gohsen, Zahra Abbasiantaeb, Mohammad Aliannejadi, Krisztian Balog, Timo Breuer, Jeffrey Dalton, Maik Fröbe, Christin Katharina Kreutz, Andreas Kruff, Simon Lupart, and 1 others. 2025. User simulation in practice: Lessons learned from three shared tasks. In *ACM SIGIR Forum*, volume 59, pages 1–18. ACM New York, NY, USA.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kalervo Järvelin and Jaana Kekäläinen. 2000. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 41–48, New York, NY, USA. Association for Computing Machinery.

Bowen Jin, Jinsung Yoon, Priyanka Kargupta, Sercan O Arik, and Jiawei Han. 2025a. An empirical study on reinforcement learning for reasoning-search interleaved llm agents. *arXiv preprint arXiv:2505.15117*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025b. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. 1996. Reinforcement learning: a survey. *J. Artif. Int. Res.*, 4(1):237–285.

Wenjun Li, Zhi Chen, Jingru Lin, Hannan Cao, Wei Han, Sheng Liang, Zhi Zhang, Kuicai Dong, Dexun Li, Chen Zhang, and 1 others. 2025. Reinforcement learning foundations for deep research systems: A survey. *arXiv preprint arXiv:2509.06733*.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Chatqa: surpassing gpt-4 on conversational qa and rag. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025a. Disco: Llm knowledge distillation for efficient sparse retrieval in conversational search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 9–19.

Simon Lupart, Daniël van Dijk, Eric Langezaal, Ian van Dort, and Mohammad Aliannejadi. 2025b. Investigating llm variability in personalized conversational information retrieval. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 353–363.

Kelong Mao, Chenlong Deng, Haonan Chen, Fengran Mo, Zheng Liu, Tetsuya Sakai, and Zhicheng Dou. 2024. ChatRetriever: Adapting large language models for generalized and robust conversational dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1227–1240, Miami, Florida, USA. Association for Computational Linguistics.

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N Angelopoulos, Trevor Darrell, Narges Norouzi, and 1 others. 2025. Search arena: Analyzing search-augmented llms. *arXiv preprint arXiv:2506.05334*.

Fengran Mo, Yifan Gao, Chuan Meng, Xin Liu, Zhuofeng Wu, Kelong Mao, Zhengyang Wang, Pei Chen, Zheng Li, Xian Li, Bing Yin, and Meng Jiang. 2025a. UniConv: Unifying retrieval and response generation for large language models in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6936–6949, Vienna, Austria. Association for Computational Linguistics.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2025b. A survey of conversational search. *ACM Trans. Inf. Syst.*, 43(6).

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 539–548, New York, NY, USA. Association for Computing Machinery.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 117–126, New York, NY, USA. Association for Computing Machinery.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. BERGEN: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics:*

EMNLP 2024, pages 7640–7663, Miami, Florida, USA. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessí, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: language models can teach themselves to use tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. 2025. Agentic reasoning and tool integration for llms via reinforcement learning. *arXiv preprint arXiv:2505.01441*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. 2025. Stepsearch: Igniting llms search ability via stepwise proximal policy optimization. *arXiv preprint arXiv:2505.15107*.

Zeqiu Wu, Yi Luan, Hannah Rashkin, D. Reitter, and Gaurav Singh Tomar. 2021. Conqrr: Conversational query rewriting for retrieval with reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing*.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Inscit: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. *arXiv preprint arXiv:2407.14482*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval*, pages 829–838.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, and 1 others. 2025. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.

Changtai Zhu, Siyin Wang, Ruijun Feng, Kai Song, and Xipeng Qiu. 2025. Convsearch-r1: Enhancing query reformulation for conversational search with reasoning via reinforcement learning. *arXiv preprint arXiv:2505.15776*.

Table 9: Prompt used for LLM-as-judge evaluation.

```
Given a Question and its Golden Answer,
verify whether the Predicted Answer is
correct. The prediction is correct if it
fully aligns with the meaning and key
information of the Golden Answer. Respond
with True if the prediction is correct and
False otherwise.

Question: {question}
Golden Answer: {golden_answer}
Predicted Answer: {predicted_answer}
```

Table 10: Rewrite prompt.

```
# Instruction:
You are given a multi-turn conversation
between a user and a system.
Rewrite the last user question into a
fully self-contained, context-independent
query.
# Guidelines:
- Include all necessary information from
previous turns.
- Resolve pronouns and vague references
into explicit entities.
- Do NOT add information not mentioned in
the conversation.
- Be concise and natural.
- Return only the rewritten query.

# Conversation:{ctx}
# Last user question:{user_utterance}
# Rewritten query:
```

A Prompts

We provide the prompts we used in this paper below. The main prompt used during the training of ChatR1 is provided in the main paper, in Table 1. We provide here the LLM-as-judge prompt in Table 9, we follow existing evaluation literature (Song et al., 2025; Rau et al., 2024). We also provide the prompt we used to generate query rewrite both with GPT-4.1 and Qwen in Table 10.

B Datasets Statistics

We provide in Table 11 the size of the five datasets we used, and their training/test splits. All datasets are composed of conversations with several turns. Datasets like TopiOCQA have long user-system interactions, with an average of more than 10 turns, while FaithDial usually has shorter interactions of around 5 turns. A turn here is defined as a pair (q,y) of both the user question and system response, making the history grow significantly.

Dataset	# Conv.	# Turns	Answer length	Corpus size
TopiOCQA	3,509 205	45,450 2,514	10.8 11.2	25M
QReCC	10,823 2,775	63,501 16,451	16.8 17.6	54M
INSCIT	249 468	1,443 2,767	30.9 40.3	49M
MultiDoc2Dial	3,469 660	18,318 3,266	20.9 20.0	10K
FaithDial	4,094 791	18,357 3,539	17.0 17.6	21M

Table 11: Statistics of the datasets used in our experiments. First and second rows correspond to train and test splits.

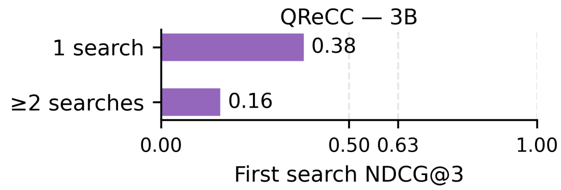

Figure 9: Performance of ChatR1 first search, and decision. (1 search): ChatR1 is satisfied with the retrieved passages and generates an answer. ($\geq$ 2 search): ChatR1 was not satisfied with the first search, so it decided to perform another search.

Since INSCIT contains both clarification turns and direct answer turns, we filtered out clarification turns. This reduces the test set from 2767 to 2197 turns. This dataset also contains turns with several gold answers for which we computed max F1 for a given turn on the two references. We filtered the training set in the same way, keeping only direct answers, and considered the multiple gold answer references as independent training samples, resulting in 1,844 samples.

We also include the corpus size. In particular, QReCC has a very large corpus, seven times larger than a dataset like MS MARCO commonly used in information retrieval. In contrast, MultiDoc2Dial is domain-specific, with a very dense passage collection on a sub-topic.

C Additional Analysis

Figure 9 provides insight into when ChatR1 decides to do another round of search. In particular, we observe that ChatR1 decides to perform another search when the first search was not satisfactory. Internally, ChatR1 thus has a threshold or its own judgment of whether the results are enough or if another search is needed.

D Baselines, and Experimental Details

We conducted our experiments with an adapted version of the Search-R1 GitHub repository[3] (Jin et al., 2025b), built upon veRL[4].

We trained several ChatR1 models, each with grounding from a different collection. When doing transfer and out-of-domain evaluation, we changed the base retrieval index to the new collection, both for baselines and ChatR1, making sure that relevant information can be found in the collection.

For the SFT baseline, we used a batch size of 512, a micro batch size of 16, a max input length of 1024, and a maximum generation length of 128 tokens. Learning rate is set to 1e-4, trained for 500 steps, with early stopping based on validation loss. This follows the setup of ChatR1.

We rely on `Qwen2.5-3B-Instruct` and its 7B variant as the base LLMs, fine-tuned on each five datasets. Retrieval model is `intfloat/e5-base-v2` used in zero-shot. ChatR1 and other baselines use the top-3 retrieved passages; we also limit the number of search calls to two, following previous works (Jin et al., 2025b,a). Query rewrites are human annotations; when not available we used GPT-4.1 to generate them. The batch size of the policy model is 512, with a PPO micro batch size of 64, a max prompt length of 3500 tokens, and a learning rate of 1e-6 for the actor model. We rely on the GAE algorithm for the critic optimization. Both policy and critic models are initialized with the same LLM, with weights fine-tuned independently. We train for 500 steps, with a saved checkpoint every 50 training steps, taking the last checkpoint by default except if training rewards collapse. SFT models are also trained in the same fashion. All experiments are conducted on 4 H100 GPUs.

E Proximal Policy Optimization (PPO)

We train ChatR1 with PPO, using the actor as policy model and critic as surrogate model. Previous works demonstrated that PPO was more stable than GRPO, hence we kept this formulation (Jin et al., 2025b). We also use a discount factor γ and GAE parameter λ, both equal to 1, following a previous study. In this setting, since the reward is only provided at the end of the trajectory, the advantage has a simplified formulation.

$$\hat{A}_i = R(\tau) - V_\psi(\tau_i) \tag{9}$$

where V_ψ is the learned value function (critic) and $R(\tau)$ the reward containing both intent and answer components. With this formulation, PPO is closer to the REINFORCE algorithm used in RL. For PPO, we also use a $\epsilon = 0.2$ in the clip loss objective.

We further provide the loss of the critic, trained with GAE in our case scenario.

$$\mathcal{L}_V(\psi) = \frac{1}{2}\,\mathbb{E}_{(q,\mathcal{H};\mathcal{R};i)\sim\mu}\left[\left(V_\psi(\tau_i) - R(\tau)\right)^2\right] \tag{10}$$

F Qualitative Examples

We provide below several qualitative examples from ChatR1.

[3] https://github.com/PeterGriffinJin/Search-R1
[4] https://github.com/volcengine/verl

Conversation History (context_block):

User: The most relaxing color to me is green, what about you?
Assistant: I don't really find any relaxing but emeralds are green due to their chromium content, do you find that interesting?
User: That is pretty interesting, I have noticed a lot of animals are green too, why is that?
Assistant: Many other creatures would beg to differ, a lot of them have adapted to their green environments by taking on their own green color to their skin.
User: Yeah that is true but I was thinking of birds and ocean creatures if any of them can see color.
Assistant: Well, I don't really know about those animals being able to see colors but green is found between yellow and blue on the visible spectrum.

User (last_user_utterance): Since green is so common they're from plants right? What makes the color green anyway?

Golden Answer:

Yes, chlorophyll is the largest contributor for green. Chlorophyll is a chemical plants use to turn sunlight into chemical energy.

Reasoning, Search, and Information:

<think> The user is asking about the presence of green in nature and what causes the color green. </think>
<search> What makes the color green? </search>
<information> Green is the color between blue and yellow on the visible spectrum. [...]</information>
<search> Why is green common in nature? </search>
<information> Green is common in nature, especially in plants. Many plants are green mainly because of a complex chemical known as chlorophyll which is involved in photosynthesis. [...]</information>
<think> From the retrieved information, green's prevalence in nature is mainly due to chlorophyll's role in photosynthesis. </think>

Predicted Answer:

<answer> I found that the color green is common in nature primarily due to the presence of chlorophyll in plants, which is involved in photosynthesis. </answer>

F1 Score: 0.300

Table 12: FaithDial (ChatR1 7B) – Full reasoning with <think>, showing how the model refines its second search query to then generate a correct answer.

Conversation History (context_block):

User: Why would I need to provide required verification for my FAFSA application?
Assistant: If you have received a note on your SAR saying you have been selected for verification or your school contacted you then you just have to confirm that the data reported on the FAFSA form is accurate. Have you been selected for verification?
User: I have not been notified that I need verification.
Assistant: Unfortunately, no relevant information is found.

User: What if I decide I want to get a loan to pay back?
Assistant: First, decide what aid you really need, and then respond to the school aid offer or choose to get a loan to pay back.

User (last_user_utterance): Okay. Let's talk about Federal Student Aid renewal. In a presumed scenario, how will I handle the renewal?

Golden Answer:

No problem. Just Login into your account and select FAFSA renewal once there. It all will develop automatically.

Reasoning, Search, and Information:

<think> The user is asking about the process of renewing Federal Student Aid (FAFSA). Typically, students need to reapply for financial aid each year, and there is a renewal process. </think>
<search> FAFSA renewal process </search>
<information> When you fill out the FAFSA form, you are applying for aid for a specific year. In order to receive aid the next year, you'll need to submit that next year's FAFSA form. [...]</information>
<search> how will I handle the renewal of Federal Student Aid </search>
<information> If you filled out a FAFSA form last year and want to renew it, select LOG IN on the FAFSA home page, select I am the student, enter your FSA ID, and be sure to select FAFSA RENEWAL once given the option. [...]</information>
<think> From the information, it is clear that to renew the FAFSA, you need to log in and select FAFSA RENEWAL. </think>

Predicted Answer:

<answer> To renew your FAFSA, you just need to log in and select FAFSA RENEWAL. This will prefill many of the nonfinancial questions for you. </answer>

F1 Score: 0.342

Table 13: MD2D (ChatR1 7B) – Reasoning on domain-specific user question, requiring clear grounding from the corpus collection.

Conversation History (`context_block`):

User: What was the song *The Night Chicago Died* about?
Assistant: In the song *The Night Chicago Died*, the narrator retells his mother's anguish while awaiting news of the fate of her husband, a Chicago policeman.
User: Who is the song *The Night Chicago Died* by?
Assistant: *The Night Chicago Died* is a song by the British group Paper Lace.

User (`last_user_utterance`): Does the group have other hits?

Golden Answer:

Billy Don't Be a Hero is a 1974 pop song that was first a UK hit for Paper Lace.

Reasoning, Search, and Information:

`<think>` Let's find out if the group Paper Lace has any other hits. =search> does the group Paper Lace have other hits? </search> My previous action is invalid. If I want to search, I should put the query between <search> and </search>. If I want to give the final answer, I should put the answer between <answer> and </answer>. Let me try again. `</think>`
`<search>` does the group Paper Lace have other hits? `</search>`
`<information>` Their first two singles, released in 1974, were both written by Murray and Callander: "Billy Don't Be a Hero" (No. 1 UK) and "The Night Chicago Died" (No. 1 US). [...] `</information>`
`<think>` Paper Lace had a hit with "Billy Don't Be a Hero". `</think>`

Predicted Answer:

`<answer>` Paper Lace had a hit with "Billy Don't Be a Hero". `</answer>`

F1 Score: 0.56

Table 14: QReCC (ChatR1 7B) – Example showing iterative self-correction during inference. While the system had to correct itself during the reasoning path, it still produced a good answer to the user.

Conversation History (context_block):

User: What type of meat is consumed at Christmas dinner in British Colonies?
Assistant: In many former British colonies the meats served are similar to those used in English Christmas dinners, and often involve roast meats. The primary meat served at Christmas in the U.K. is turkey, though historically boar, peacock, swan, and goose have also been served.
User: What are the food served at Christmas dinner in Japan?
Assistant: A successful advertising campaign in the 1970s made eating at KFC around Christmas a national custom. Its chicken meals are so popular during the season that stores take reservations months in advance.
User: How about Christmas dinner in Lebanon? What kind of food do they eat?
Assistant: In Lebanon, roast turkey is the most common choice of meal. Roasted duck, Lebanese salad (Tabbouleh), and pastries, such as honey cake, are common.

Query (last_user_utterance): Now I want to know the Christmas dinner in Honduras!

Golden Answer:

Christmas dinner is served around midnight on the 24th of December, consisting of tamales, roast pork leg, accompanied by "torrejas," for dessert, and eggnog.

Search and Information:

<search> Now I want to know the Christmas dinner tradition in Honduras? </search>
<information> The traditional Honduran Christmas dinner is served around midnight on the 24th of December. [...]</information>
<search> What kind of food is typically eaten during Christmas dinner in Honduras? </search>
<information> In Honduras, tamales are traditionally eaten on Christmas Eve evening. [...]</information>

Predicted Answer:

<answer> In Honduras, the traditional Christmas dinner is served around midnight on the 24th of December and typically consists of tamales, roast pork leg, accompanied by "torrejas," for dessert, and eggnog. </answer>

F1 Score: 0.8627

Table 15: INSCIT (ChatR1 3B) – Example where the model infers from the context that the user is interested in examples of food, it thus refines the second search to better adapt to the user intent. We also see reduced thinking process here, reasoning is implicit, and made through the search queries.